

Training Space Truncation in Vision-Based Recognition

René Dencker Eriksen and Ivar Balslev

University of Southern Denmark, Odense,
The MaerskMc-Kinney Møller
Institute for Production Technology

Abstract. We report on a method for achieving a significant truncation of the training space necessary for recognizing rigid 3D objects from perspective images. Considering objects lying on a table, the configuration space of continuous coordinates is three-dimensional. In addition the objects have a few distinct support modes. We show that recognition using a stationary camera can be carried out by training each object class and support mode in a two-dimensional configuration space. We have developed a transformation used during recognition for projecting the image information into the truncated configuration space of the training. The new concept gives full flexibility concerning the position of the camera since perspective effects are treated exactly. The concept has been tested using 2D object silhouettes as image property and central moments as image descriptors. High recognition speed and robust performance are obtained.

Keywords: Computer vision for flexible grasping, recognition of 3D objects, pose estimation of rigid object, recognition from perspective images, robot-vision systems

1 Introduction

We describe here a method suitable for computer-vision-based flexible grasping by robots. We consider situations where classes of objects with known shapes, but unknown position and orientation are to be classified and grasped in a structured way. Such systems has many quality measures such as recognition speed, accuracy of the pose estimation, low complexity of training, free choice of camera position, generality of object shapes, ability to recognize occluded objects, and robustness. We shall evaluate the properties of the present method in terms of these quality parameters.

Recognition of 3D objects has been widely based on establishing correspondences between 2D features and the corresponding features on the 3D object [1-3]. The features has been point like, straight lines or curved image elements. The subsequent use of geometric invariants makes it possible to classify and pose estimate objects [4-8]. Another strategy is the analysis of silhouettes. When perspective effects can be ignored, as when the objects are flat and the camera is

remote, numerous well established methods can be employed in the search for match between descriptors of recorded silhouette and those of silhouettes in a data base [9-11]. Other method are based on stereo vision or structured light [1, 11-12].

In the present paper we face the following situation:

- The rigid objects do not have visual features suitable for 2D-3D correspondence search or 2D-2D correspondence search used in stereo vision.
- No structured light is employed.
- The camera is not necessarily remote, so that we must take perspective effects into account.

We propose here a 'brute force' method [13] in which a large number of images or image descriptors are recorded or constructed during training. Classification and pose estimation is then based on a match search using the training data base. A reduction of the configuration space of the training data base is desirable since it gives a simpler training process and smaller extent of the data bases. The novelty of the present method is the recognition based on training in a truncated configuration space.

The method is based on a nonlinear 2D transformation of the image to be recognized. The transformation corresponds to a virtual displacement of the object into an already trained position relative to the camera. As relevant for many applications we consider objects lying at rest on a table or conveyer belt. In Sect. 2 we describe the 3D geometry of the system. We introduce the concept, 'virtual displacement', and define the truncated training space. In Sect. 3 are described the relevant 2D transformation and the match search leading to the classification and pose estimation. We also specify our choice of descriptors and match criterion in the recognition. The method has been implemented by constructing a physical training setup and by developing the necessary software for training results and recognition. Typical data of the setup and the objects tested are given in Sect. 4. We also present a few representative test results.

The work does not touch upon the 2D segmentation on which the present method must rely. The 2D segmentation is known to be a severe bottleneck if the scene illumination and relative positions of the objects are inappropriate. In the test we used back-lighting and nonoccluded objects in order to avoid such problems. Therefore we can not evaluate the system properties in case of complex 2D segmentation.

2 The 3D Geometry and the Truncated Training Space

In the present work we consider a selection of physical objects placed on a horizontal table or a conveyer belt, see Fig.1. The plane of the table surface is denoted π . We consider gravity and assume object structures having a discrete number of ways - here called support modes - on which the surface points touch the table. This means that we exclude objects, which are able to roll with a constant height of the center-of-mass. Let i count the object classes and let j

count the support modes. With fixed j , each object's pose has three degrees of freedom, e.g. (x, y, ω) , where (x, y) is the projection on the table plane π of its center-of-mass (assuming uniform mass density), and ω is the rotation angle of the object about a vertical axis through the center-of-mass.

A scene with one or more objects placed on the table is viewed by an ideal camera with focal length f , pin hole position H at a distance h above the table, and an optical axis making an angle α with the normal to the plane π , see Fig. 1. The origin of (x, y) is H 's projection O on π , and the y -axis is the projection of the optical axis on π . Let (x, y, z) be the coordinates of a reference point of the object. We introduce the angle ϕ defined by:

$$\cos \phi = \frac{y}{\sqrt{x^2 + y^2}}, \quad \sin \phi = \frac{x}{\sqrt{x^2 + y^2}}. \quad (1)$$

Consider a virtual displacement of the object so that its new position is given by:

$$(x', y', \omega') = (0, \sqrt{x^2 + y^2}, \omega - \phi) \quad (2)$$

This displacement is a rotation about a vertical axis through the pinhole. Note that the same same points on the object surface are visible from the pin hole H in the original and displaced position. The inverse transformation to be used later is given by:

$$x = y' \sin \phi, \quad y = y' \cos \phi, \quad \omega = \omega' + \phi \quad (3)$$

The essential property of this transformation is that the corresponding 2D transformation is independent of the structure of the object. The truncation of the training space introduced in the present paper is based on this property.

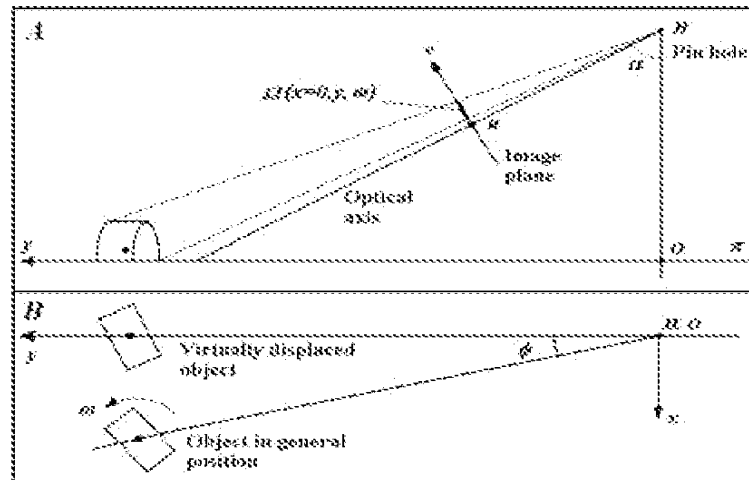


Fig. 1. Horizontal (A) and vertical (B) views of the system including the table, the camera, and the object before and after the virtual displacement.

We focus on image properties condensed in binary silhouettes. Therefore, we assume that the scene is arranged with a distinguishable background color so

that each object forms a well defined silhouette $\Omega(i, j, x, y, \omega)$ on the camera image. Thus, $\Omega(i, j, x, y, \omega)$ is a list of coordinates (u, v) of set pixels in the image. We assume throughout that $(u, v) = (0, 0)$ is lying on the optical axis. The task in the present project is to determine (i, j, x, y, ω) from a measurement of an object's silhouette Ω_o and a subsequent comparison with the silhouettes $\Omega(i, j, x = 0, y, \omega)$ recorded or constructed in a reduced configuration space. In the data base the variables y and ω are suitably discretized. Silhouettes for the data base are either recorded by a camera using physical objects or constructed from a CAD representation.

3 The 2D Transformation and the Match Search

After the above mentioned virtual 3D displacement, an image point (u, v) of the object will have the image coordinates (u', v') given by:

$$u'(\phi, u, v) = \frac{f(u \cos \phi + v \sin \phi \cos \alpha - f \sin \phi \sin \alpha)}{u \sin \phi \sin \alpha + v(1 - \cos \phi) \sin \alpha \cos \alpha + f(\cos \phi \sin^2 \alpha + \cos^2 \alpha)} \quad (4)$$

$$v'(\phi, u, v) = \frac{f(-u \sin \phi \cos \alpha + v(\cos \phi \cos^2 \alpha + \sin^2 \alpha) + f(1 - \cos \phi) \sin \alpha \cos \alpha)}{u \sin \phi \sin \alpha + v(1 - \cos \phi) \sin \alpha \cos \alpha + f(\cos \phi \sin^2 \alpha + \cos^2 \alpha)} \quad (5)$$

This result can be derived by considering - in stead of an object displacement - three camera rotations about H : A tilt of angle $-\alpha$, a roll of angle ϕ , and a tilt of angle α . Then the relative object-camera-position is the same as if the object were displaced according to the 3D transformation described in Sect. 2. Note that the inverse transformation corresponds to a sign change of ϕ .

By transforming all points in a silhouette Ω according to (4-5), one obtains a new silhouette Ω' . Let us denote this silhouette transformation T_ϕ , so that

$$\Omega' = T_\phi(\Omega) \quad (6)$$

The 2D center-of-mass of Ω is $(u_{cm}(\Omega), v_{cm}(\Omega))$. The center-of-mass of the displaced silhouette Ω' is close to the transformed center-of-mass of the original silhouette Ω . In other words

$$u_{cm}(\Omega') \approx u'(\phi, u_{cm}(\Omega), v_{cm}(\Omega)) \quad (7)$$

$$v_{cm}(\Omega') \approx v'(\phi, u_{cm}(\Omega), v_{cm}(\Omega)) \quad (8)$$

This holds only approximately because the 2D transformation in (4-5) is nonlinear. In Fig 2 is shown a situation in which Ω' is a square. The black dot in Ω' is the transformed center-of-mass of Ω , which is displaced slightly from the center-of-mass of Ω' .

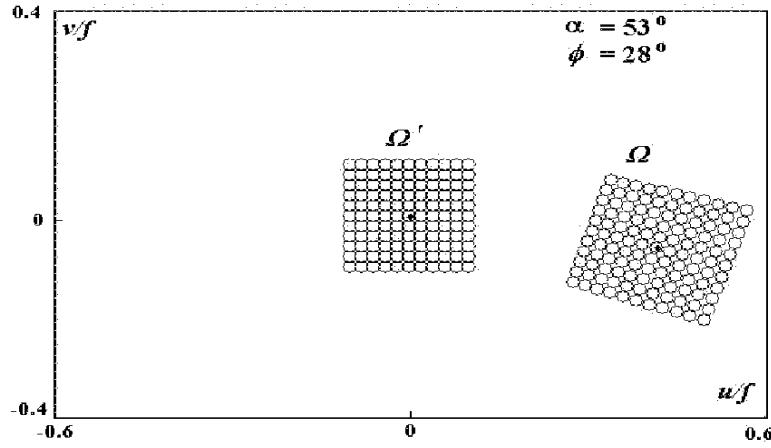


Fig. 2. An image component Ω and the transformed version Ω' in case that Ω' is a square. The values of α and ϕ are given in the upper right corner. The black dot Ω is the 2D center-of-mass. After transformation this point has a position shown as the black dot in Ω' .

Let $\Omega_{tr} = \Omega_{tr}(i, j, y, \omega)$ be the silhouettes of the training with $x = 0$. The training data base consist of descriptors of $\Omega_{tr}(i, j, y, \omega)$ with suitably discretized y and ω . In case of not too complex objects,

$$u_{cm}(\Omega_{tr}) \approx 0 \quad (9)$$

The object to be recognized has the silhouette Ω_o . This silhouette defines an angle ϕ_o given by

$$\phi_o = \arctan\left(\frac{u_{cm}(\Omega_o)}{f \sin \alpha - v_{cm}(\Omega_o) \cos \alpha}\right) \quad (10)$$

According to (4, 7, 10), the transformed silhouette $\Omega'_o = T_{\phi_o}(\Omega_o)$ has a 2D center-of-mass close to $u = 0$:

$$u_{cm}(\Omega'_o) \approx 0 \quad (11)$$

We shall return to the approximations (7-9,11) later.

Eqs. (9) and (11) imply that Ω'_o is to be found among the silhouettes $\Omega_{tr}(i, j, y, \omega)$ of the data base. Because of the approximations (9,11), the similarity between Ω'_o and $\Omega_{tr}(i, j, y, \omega)$ is not exact with regards to translation, so one must use translational invariant descriptors in the comparison.

In the search for match between $\Omega_{tr}(i, j, y, \omega)$ and Ω'_o it is convenient to use that $v_{cm}(\Omega'_o) \approx v_{cm}(\Omega'_{tr}(i, j, y, \omega))$. It turns out, that $v_{cm}(\Omega'_{tr}(i, j, y, \omega))$ is usually a monotonous function of y , so - using interpolation - one can calculate a data base slice with a specified value v_{cm} and with i, j, ω as entries. This means that i, j , and ω can be determined by a match search between moments of Ω'_o

and moments in this data base slice. Note that the data base slice involves one continuous variable ω . With a typical step size of 3° the data base slice has only 120 records per support mode and per object class.

The result of the search are i_{match} , j_{match} , and ω_{match} . The value y_{match} can be calculated using the relation between y and $v_{cm}(\Omega_{tr})$ for the relevant values of i , j , and ω . The original pose (x, y, ω) can now be found by inserting $y' = y_{\text{match}}$, $\omega' = \omega_{\text{match}}$, and $\phi = \phi_o$ in Eq. (3).

If the approximation (9) brakes down, one must transform all the silhouettes of the data base, so that the match search takes place between Ω'_o and $\Omega'_{tr} = T_{\phi_{tr}}(\Omega_{tr})$ where

$$\phi_{tr} = \arctan\left(\frac{u_{cm}(\Omega_{tr})}{f \sin \alpha - v_{cm}(\Omega_{tr}) \cos \alpha}\right) \quad (12)$$

In this case $\phi = \phi_o - \phi_{tr}$ should be inserted in (3) in stead of ϕ_o .

We are left with the approximation (11), demonstrated in Fig. 2. This gives a slightly wrong angle ϕ_o . If the corresponding errors are harmful in the pose estimation, then one must perform an iterative calculation of ϕ_o , so that (11) holds exactly.

We conclude this section by specifying our choice of 1) image descriptors used in the data base, and 2) recognition criterion. In our test we have used as descriptors the 8-12 lowest order central moments, namely μ_{00} , μ_{20} , μ_{11} , μ_{02} , μ_{30} , μ_{21} , μ_{21} , and μ_{03} . The first order moments are absent since we use translational invarianat moments. In addition we used in some of the tests, the width and height of the silhouette. In the recognition strategy we minimized the Euclidean distance in a feature space of descriptors. The descriptors were normalized in such a way that the global variance of each descriptor was equal to one [14].

4 The Experiments

Fig. 3 shows the setup for training and test. We used a rotating table for scanning through the training parameter ω and a linear displacement of the camera for scanning through the parameter y . The pose estimation was checked by a grasping robot. In order to avoid 2D segmentation problems we used backlighting in both training and recognition.

The parameters for the setup and typical objects tested are shown in the Table 1.

We report here the result of a test of a single object, a toy rabbit manufactured by LEGO[®], see Fig 4. Its 5 support modes are shown along with the support mode index used. After training using an angular step size of $\Delta\omega = 4^\circ$, we placed the rabbit with one particular support mode 100 random poses, i.e. values of (x, y, ω) in the field of view. The support modes detected by the vision system were recorded. This test was repeated for the remaining support modes. The results are shown in the Table 2. The two confused support modes were

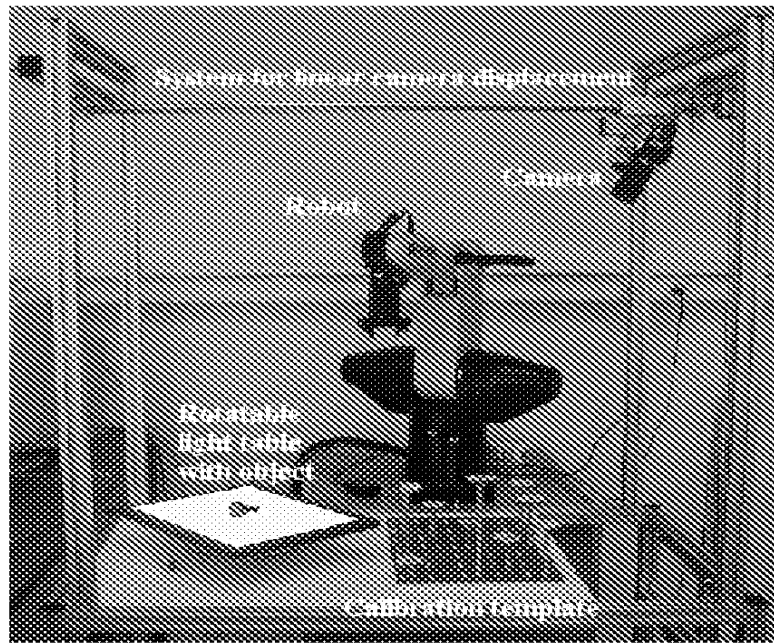


Fig. 3. Setup for training and test. The calibration template is used for calibrating the camera relative to a global coordinate system.

Table 1. Properties and parameters of the objects and the test setup.

Angle α	25°
Height h	800 mm
Field of view	400 mm x 300 mm
$\Delta\omega$ = angular step size during training	4° - 7.2°
Δy = translational step size during training	50 mm
Camera resolution (pixels)	768 x 576
Number of support modes of objects	3-5
Typical linear object dimensions	25-40 mm
Typical linear dimensions of object images	40-55 pixels
Silhouette descriptors	8 lowest order centr. moments + width & height of silhouette
Number of data base records per object	900-1500 for $\Delta\omega = 7.2^\circ$
Training time per support mode	5 min.
Recognition time (after 2D segmentation)	5-10 ms

'support by four paws' and 'support by one ear and fore paws'. It can be understood from Fig. 4, that these two support modes are most likely to be mixed up in a pose estimation. We repeated the experiment with $\Delta\omega = 7.2^\circ$. In this case no errors were measured in the 500 tests.

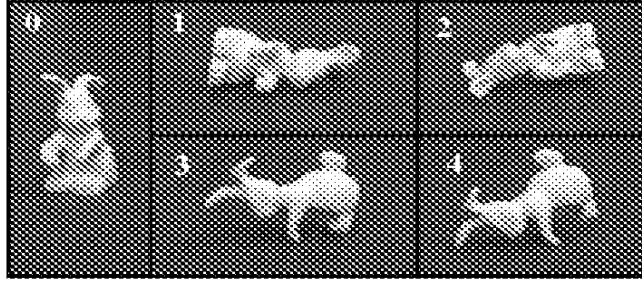


Fig. 4. The toy rabbit shown in its five support modes. The support mode indices used in Table 2 are written in the upper left corner.

Table 2. Statistics of the support mode detection when the toy rabbit was placed at random positions (x, y, ω) in the field of view. Each support mode was tested 100 times and the experiments involved two different angular step sizes in the training.

Angular step size	7.2°					4°				
True ↓, Detected →	0	1	2	3	4	0	1	2	3	4
0 standing	100					100				
1 lying on left side		100					100			
2 lying on right side			100					100		
3 on fore & hind paws				98	2				100	
4 on ear & fore paws				1	99					100

5 Discussion

A complete vision system for flexible grasping consists of two processes, one performing the 2D segmentation, and one giving the 3D interpretation. We have developed a method to be used in the second component only, since we used a illumination and object configuration giving very simple and robust segmentation.

The method developed is attractive with respect to the following aspects:

- High speed of the 3D interpretation.
- Generality concerning the object shape.
- Flexibility of camera position and object shapes, since tall objects, closely positioned cameras, and oblique viewing directions are allowed. In case of ambiguity in the pose estimation when viewed by a single camera, it is easy to use 2 or more cameras with independent 3D interpretation.
- Simple and fast training without assistance from vision experts.

The robustness and total recognition speed depends critically on the 2D segmentation, and so we can not conclude on these two quality parameters. The method in its present form is not suitable for occluded objects.

One remaining property to be discussed is the accuracy of the pose estimation. In our test the grasping uncertainty was about ± 2 mm and $\pm 3^\circ$.

However, the origin of these uncertainties were not traced, so they may be reduced significantly by careful camera-robot co-calibration.

In our experiments we used a rather coarse discretization of y and ω , and only one object at a time was recognized. The recognition time in the experiment was typically 20 ms per object (plus segmentation time). This short processing times gives plenty of room for more demanding tasks involving more objects, more support modes, and higher accuracy through a finer discretization of y and ω .

6 Conclusion

We have developed and tested a computer vision concept appropriate in a brute force method based on data bases of image descriptors. We have shown that a significant reduction of the continuous degrees of freedom necessary in the training can be achieved by applying a suitable 2D transformation during recognition prior to the match search. The advantages are the reductions of the time and the storage used in the training process.

The prototype developed will be used for studying a number of properties and possible improvements. First, various types of descriptors and classification strategies will be tested. Here, color and gray tone information should be included. Second, the over-all performance with different 2D segmentation strategies will be studied, particularly those allowing occluded objects. Finally, the concept of training space truncation should be extended to systems recognizing objects of arbitrary pose.

References

1. B.K.P. Horn, Robot Vision, The MIT Press, Cambridge, Massachusetts, 1998.
2. R.M. Haralick, L.G. Shapiro, Computer and Robot Vision, Vol II, Addison Wesley, Reading, Massachusetts, 1993.
3. I. K. Park, K.M. Lee, S. U. Lee, Recognition and Reconstruction of 3D objects using Model-based Perceptual Grouping, Proc. Int. Conf Pattern Recognition 2000, Barcelona, Vol. 1 (A. Sanfeliu et. al., eds.) pp. 720-724, IEEE Computer Society, Los Alamitos
4. K. Nagao, B.K.P. Horn, Direct Object Recognition Using Higher Than Second Order Statistics of the Images, A.I.Memo#1526, MIT, 1995.
5. K. Nagao, W.E.L. Grimson, Using Photometric Invariants for 3D Object Recognition, Computer Vision and Image Understanding 71, 1998, 74-93.
6. W.E.L. Grimson, Object Recognition by Computer: The Role of Geometric Constraints, The MIT Press, Cambridge, Massachusetts, 1990
7. J. L. Mundy and A. Zisserman, Repeated Structures: Image Correspondence Constraints and 3D Structure Recovery, in Applications of Invariance in Computer Vision Springer-Verlag, Berlin (J.L. Mundy, A. Zisserman, and D. Forsyth, eds), 1994, pp. 89-106.
8. J. Ponce, D.J. Kriegman, Toward 3D Object Recognition from Image Contours, in 'Geometric Invariance in Computer Vision', (J.L. Mundy and A. Zisserman, eds.), The MIT Press, Cambridge, Massachusetts, 1992.

9. N. Götze, S. Drüe, G. Hartmann, Invariant Object Recognition with Discriminative Features based on Local Fast-Fourier Mellin Transformation, Proc. Int. Conf. Pattern Recognition 2000, Barcelona, Vol. 1 (A. Sanfeliu et. al., eds.) pp. 948-951, IEEE Computer Society, Los Alamitos.
10. S. Abbasi, F. Mokhtarian, Automatic View Selection and Signal Processing, Proc. Int. Conf. Pattern Recognition 2000, Barcelona, Vol. 1 (A. Sanfeliu et. al., eds.) pp.13-16, IEEE Computer Society, Los Alamitos
11. R.C.G. Gonzales, R.E. Woods, Digital Image Processing, Addison Wesley, Reading, Massachusetts, 1992.
12. M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis and Computer Vision, PWS Publishing, Pacific Grove, California, 1999.
13. R.C. Nelson, A. Selinger, Learning 3D Recognition Models for General Objects from Unlabeled Imagery: An Experiment in Intelligent Brute Force, Proc. Int. Conf. Pattern Recognition 2000, Barcelona, Vol. 1 (A. Sanfeliu et. al., eds.) pp. 1-8, IEEE Computer Society, Los Alamitos
14. Balslev, Noise Tolerance of Moment Invariants in Pattern Recognition, Pattern Recognition Letters 19, 1998, 183-89.